Archival Storage for High-End Computing Systems

Ethan L. Miller

Storage Systems Research Center

University of California, Santa Cruz

Ian Adams • Kevin Greenan • Andrew Leung • Darrell Long

Brian Madden • David Rosenthal • Daniel Rosenthal • Mark Storer

Kaladhar Voruganti • Lee Ward • Avani Wildani

What is archival storage, and why do we need it?

What is archival storage, and why do we need it?

- Functional: a way of passing information to future generations
  - Science builds on earlier research
  - Much of today's knowledge is in the form of electronic data
  - Future researchers need access to this data!

What is archival storage, and why do we need it?

- Functional: a way of passing information to future generations
  - Science builds on earlier research
  - Much of today's knowledge is in the form of electronic data
  - Future researchers need access to this data!

- Operational: systems, tools and techniques for long-term storage
  - Actual computer systems
  - Data management tools
  - Techniques for interpreting data and preserving contexts
  - All of these need to evolve over time!

More motivation

- Recent NSF report:
  Advisory Committee for Cyberinfrastructure Task Force on Data and Visualization (March 2011)
- “Preserving data to preserve the planet”
- Recommended establishing scientific data archives
- Archives are necessary in most scientific disciplines
  - Only a few disciplines have organized archives today...
  - Building them to be both cost-effective and useful is an important problem!

Challenges

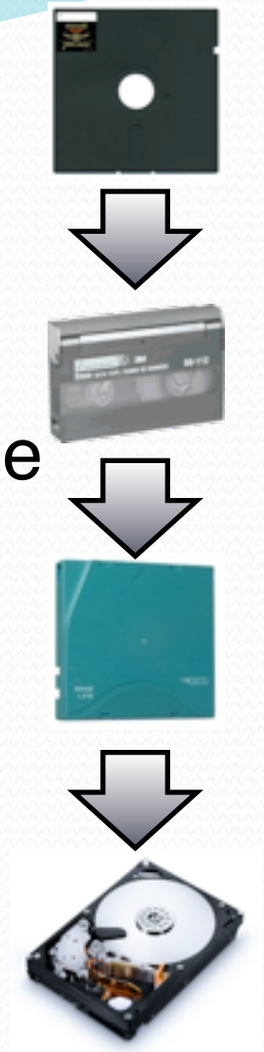
- Preserving data (bits) for a very long time
- Finding the data you're looking for
  - Includes the ability to actually retrieve it!
- Comprehending the data at a (much) later time
- **Need to do *all* of these for archival storage to be effective**

Challenge: preserving bits for the long term

- System must outlast any individual component!
  - Devices have a limited lifetime
  - Newer devices are often more efficient (space, energy)
  - Systems (hardware & software) must evolve smoothly over time
- Reliability is a crucial issue
  - Guard against data loss at all levels: device, system, site
    - Repairs should be as “localized” as possible
  - Repair loss as soon as practical
    - Trade off repair time and power consumption
- System must scale to very large scale
- System cost is a big issue: long-term cost per year of storing data
  - Acquisition cost (including system maintenance and replacement)
  - Operational cost (people, power, etc.)
  - Models must include discount rates, technology growth over time

Challenge: preserving bits for the long term

- System must outlast any individual component!
  - Devices have a limited lifetime
  - Newer devices are often more efficient (space, energy)
  - Systems (hardware & software) must evolve smoothly over time
- Reliability is a crucial issue
  - Guard against data loss at all levels: device, system, site
    - Repairs should be as “localized” as possible
  - Repair loss as soon as practical
    - Trade off repair time and power consumption
- System must scale to very large scale
- System cost is a big issue: long-term cost per year of storing data
  - Acquisition cost (including system maintenance and replacement)
  - Operational cost (people, power, etc.)
  - Models must include discount rates, technology growth over time



Challenge: finding data in archives

- Archives contain petabytes (soon, exabytes) of data
  - Users that stored the data may not be around to help find it
    - Many researchers keep notes in text files or lab notebooks
  - Researchers want to “mine” the archive for useful data
- Search is critical!
  - Need to maintain more than just basic file metadata
- Centralized databases may not work well
  - Might not scale to millions of archive devices
  - Much more vulnerable to failure and corruption
    - Difficult to recover from it!
- Resource consumption is an issue: archives can’t be profligate

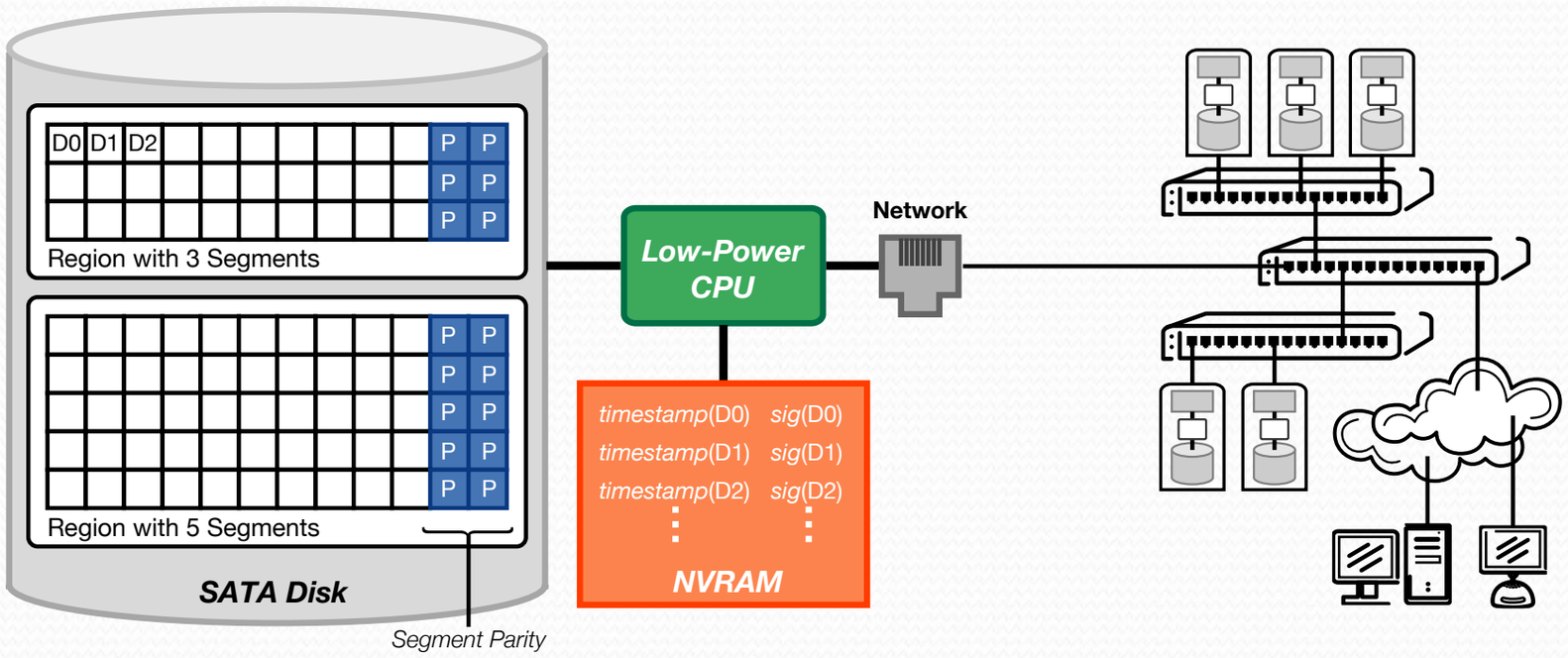
Challenge: comprehending archived data

- Preserving and finding the data is good, but we need to be able to *understand* it
  - Difficult when the software that generated it may no longer be runnable
  - Especially true for HEC applications, which may run on long-gone systems
- Standards are essential, but may not be enough
  - Preserve execution environments?
  - Better metadata explaining stored data?

Meta-challenge: understanding archival workloads

- Few current studies of long-term data usage
  - Especially in HEC environments!
- Critical for building new archives: long-held assumptions may be incorrect
  - Immutability of files
  - Access density to stored data
  - Predictability of workload
- Measuring existing systems will help us better design new archival storage systems to meet actual needs

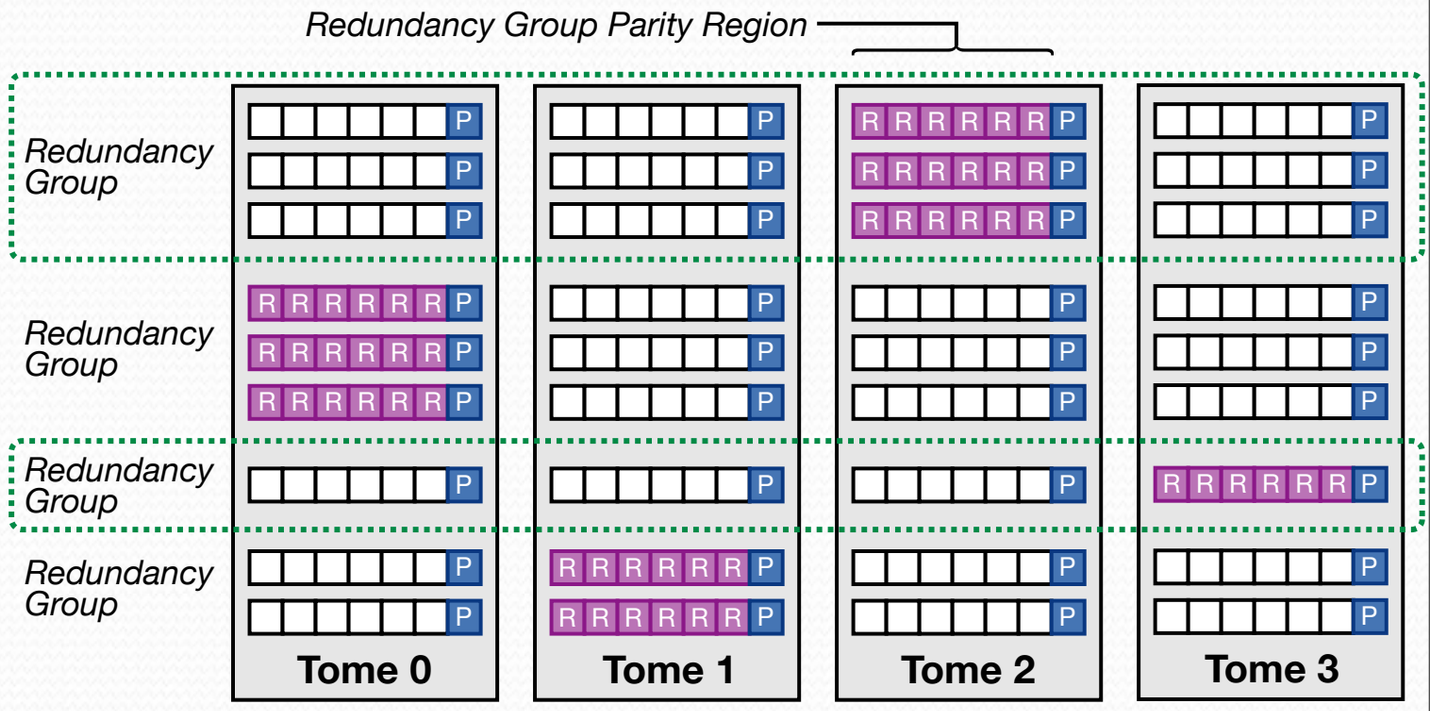
Research: Pergamum & DAWN

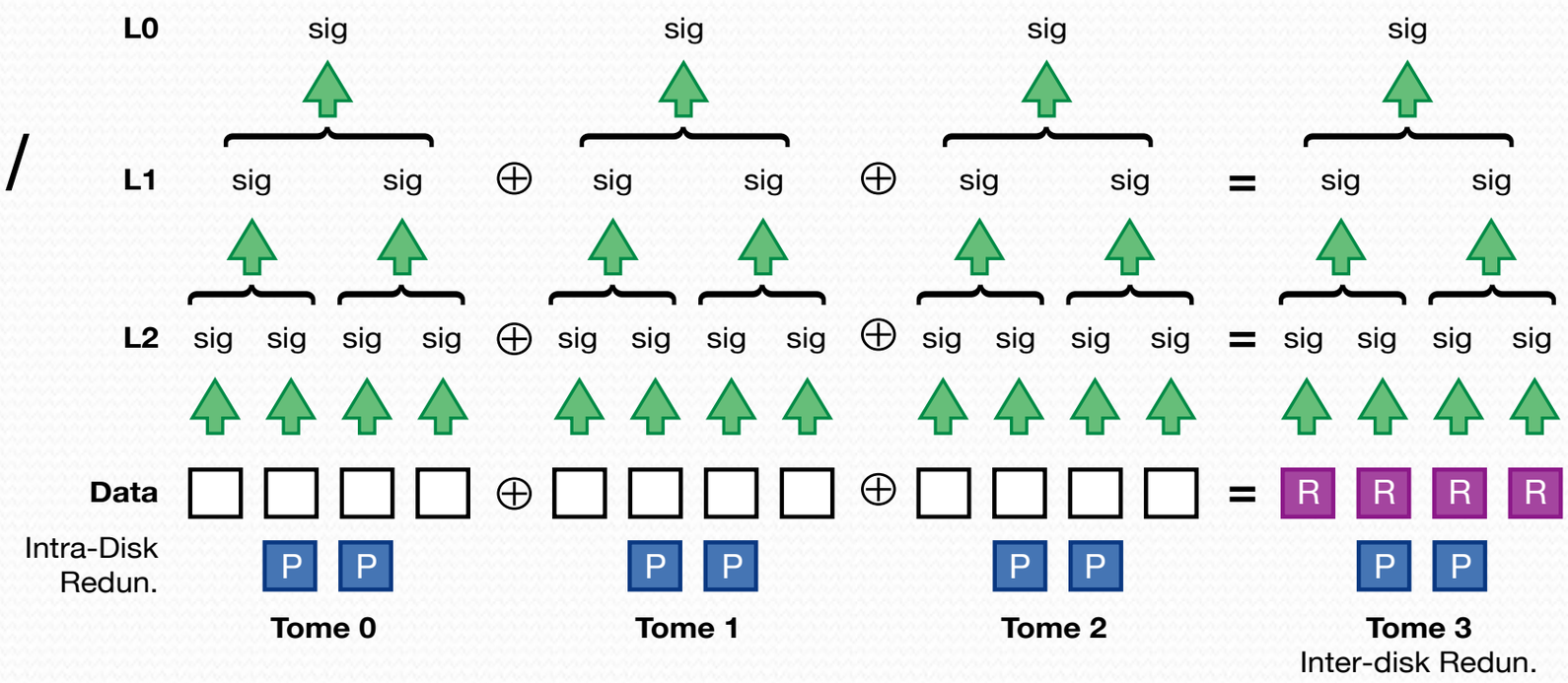


- Scalable, evolvable archival storage from smart devices
  - Pergamum: disk-based
  - DAWN: flash-based
- Smooth system evolution: based around flexible network protocols
- Scalable reliability mechanisms
- Ongoing research: scalable indexing mechanisms

Pergamum & DAWN: ensuring data reliability

- Use multi-level reliability: detect and correct errors as locally as possible
- Constantly monitor data for changes
  - Trade off detection / correction rate and power / cost





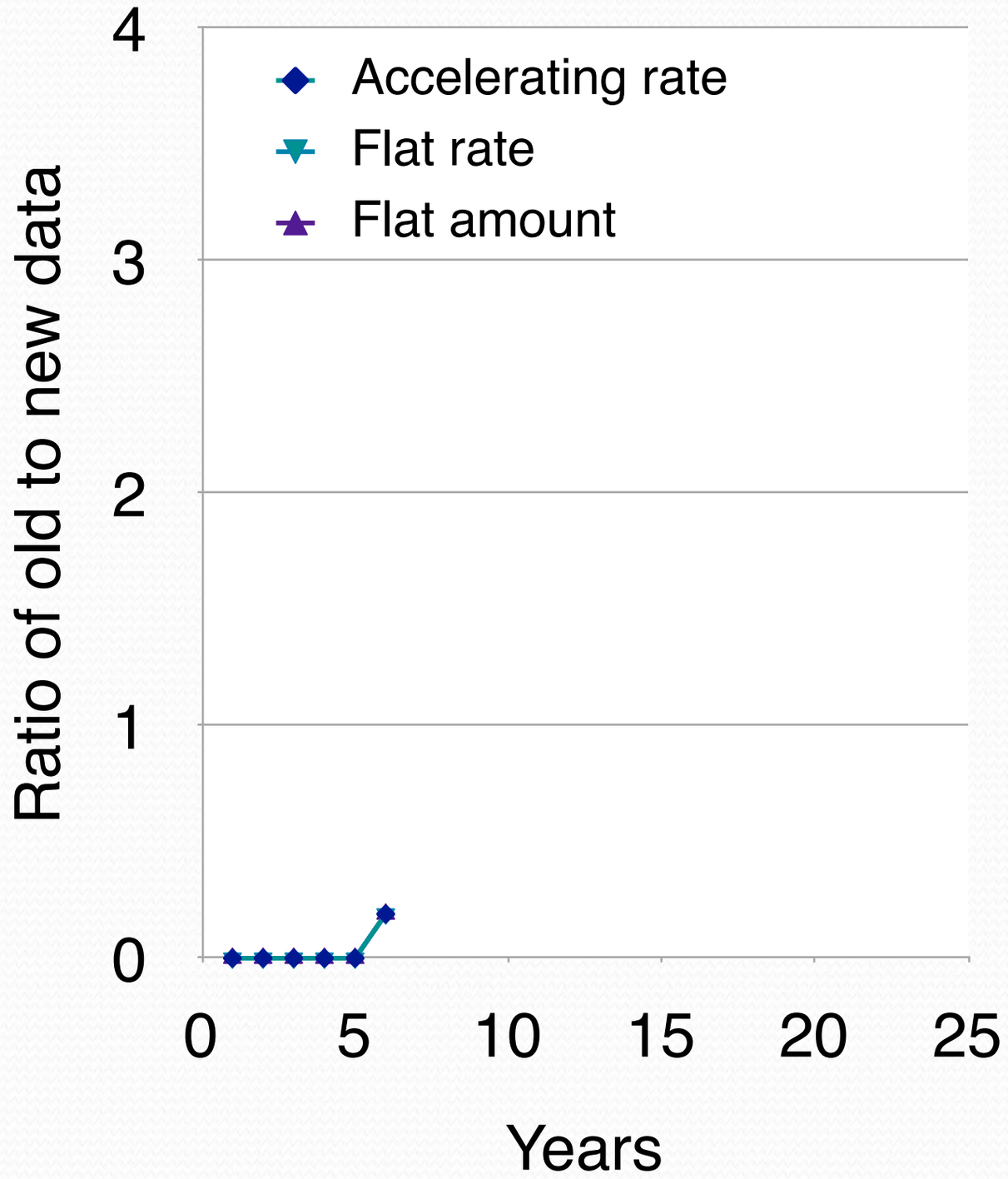
DAWN: flash for archival storage?

- Proposal: use flash for archival storage
  - Higher upfront cost
  - Lower operating cost?
    - Lasts longer
    - More reliable (and fails in different ways)
    - Lower power
- Do these costs balance?
  - At what point (if any) is flash worth it?
  - Similar questions can be asked of disk and tape...
- Goal: model economic tradeoffs inherent in building archival storage
  - More critical for archival storage: small changes in assumptions can mean large effects
  - More need for accurate forecasting

Research: economics of archival storage

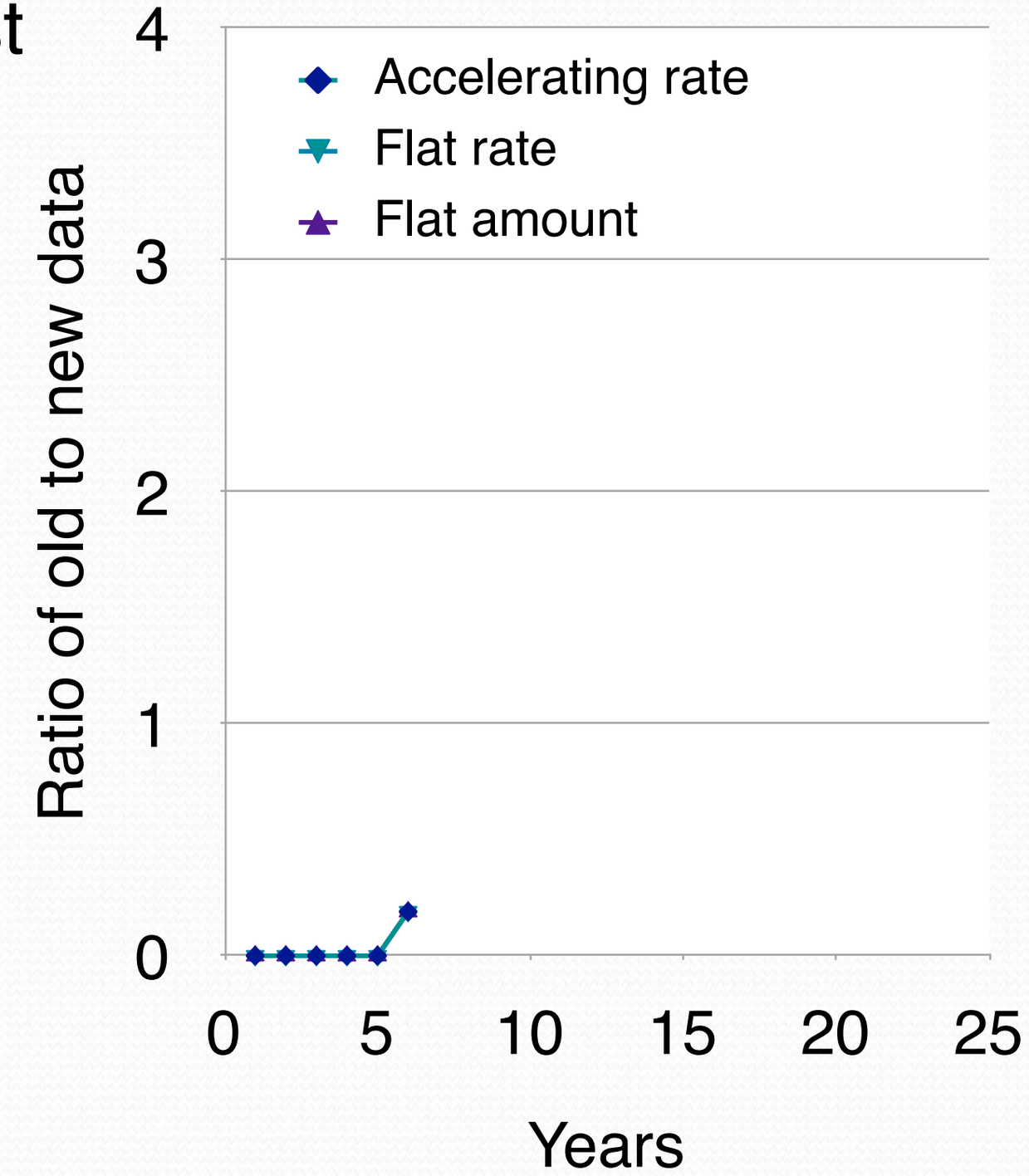
- Users want to pay for archival storage once: when data is created
  - New data is most frequently used
  - Many commercial models collect money from usage (Flickr, YouTube)
- Problem: archival storage has ongoing costs!
  - Refresh cycles for data and media
  - Management costs
- **Usage falls off dramatically as data ages!**
  - Trade off high initial cost against high ongoing costs?
    - Fewer refresh cycles & lower management cost?
  - Pay for ongoing storage with revenue from new data?
    - Depends on increasing growth rate: not sustainable in the long term
  - Get rid of much of the data
    - Which data and who decides?

Example: impact of growth models on cost



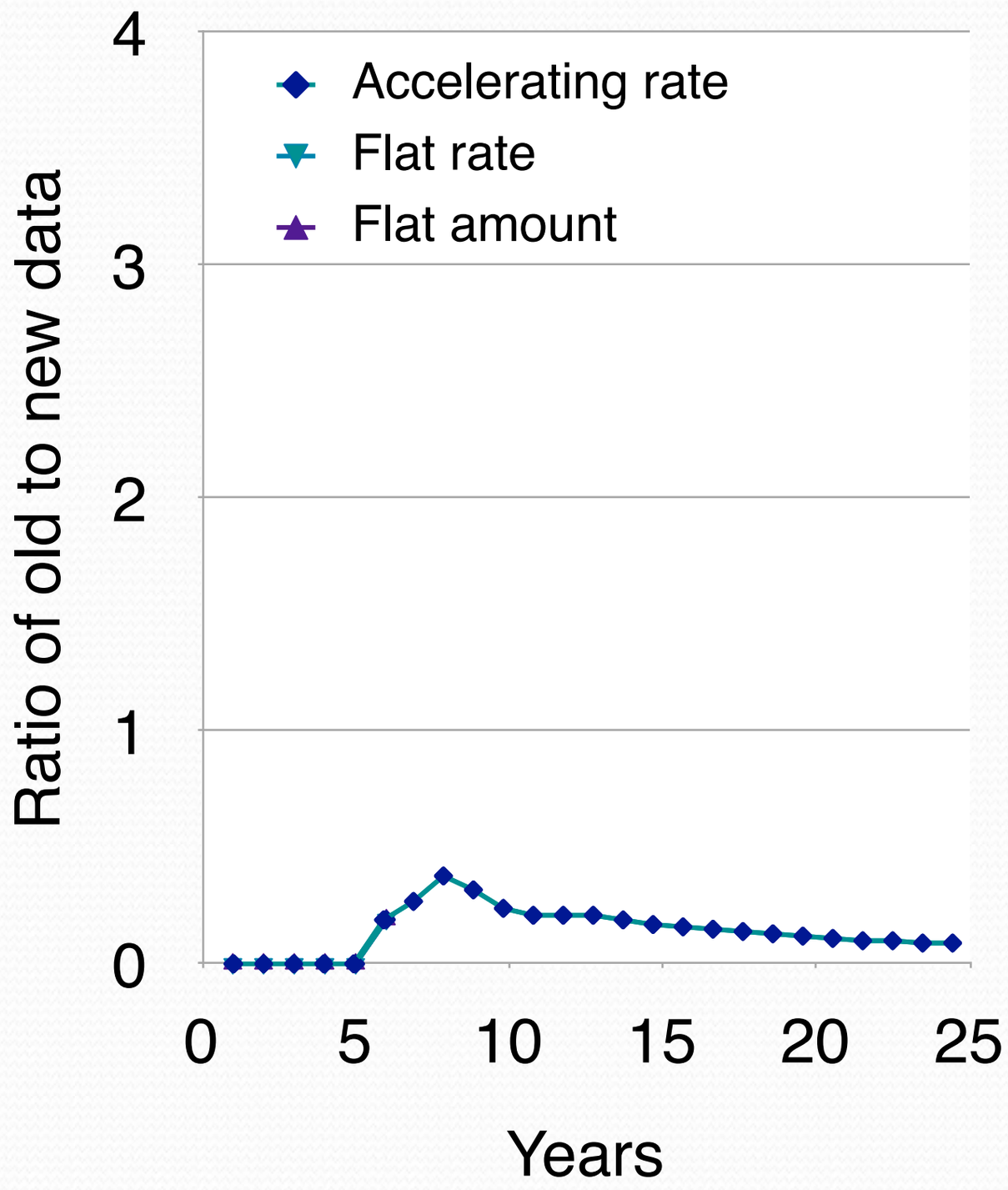
Example: impact of growth models on cost

- Exponential growth for first 5 years
  - Slows a bit in years 4–5



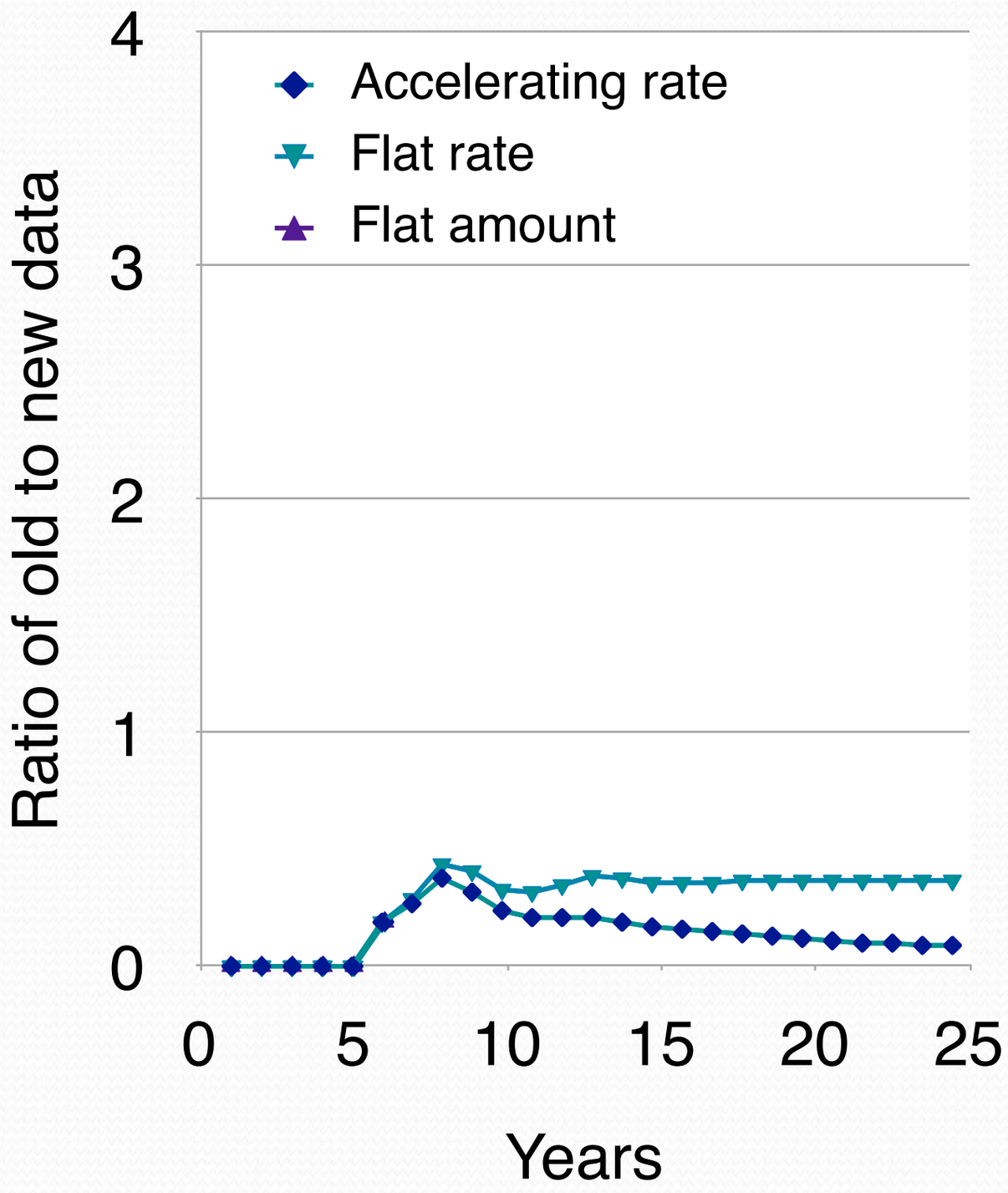
Example: impact of growth models on cost

- Exponential growth for first 5 years
  - Slows a bit in years 4–5
- Increasing growth rate
  - New storage costs dominate existing storage
  - Ratio of old:new drops over time



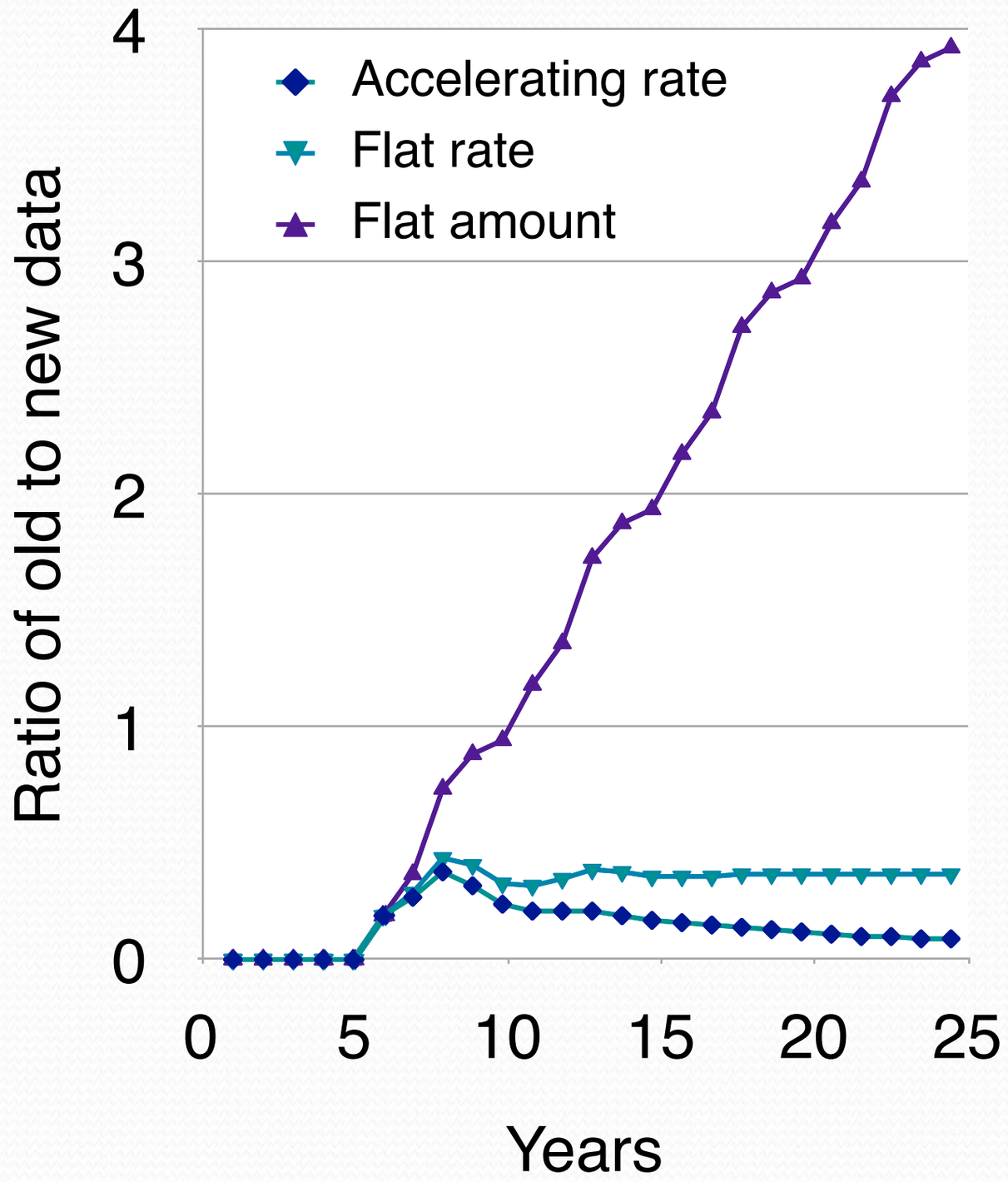
Example: impact of growth models on cost

- Exponential growth for first 5 years
  - Slows a bit in years 4–5
- Increasing growth rate
  - New storage costs dominate existing storage
  - Ratio of old:new drops over time
- Level growth rate
  - Old data : new data ratio remains approximately constant



Example: impact of growth models on cost

- Exponential growth for first 5 years
  - Slows a bit in years 4–5
- Increasing growth rate
  - New storage costs dominate existing storage
  - Ratio of old:new drops over time
- Level growth rate
  - Old data : new data ratio remains approximately constant
- Level growth *amount*
  - Old data dominates quickly



Ongoing research: understanding HEC workloads

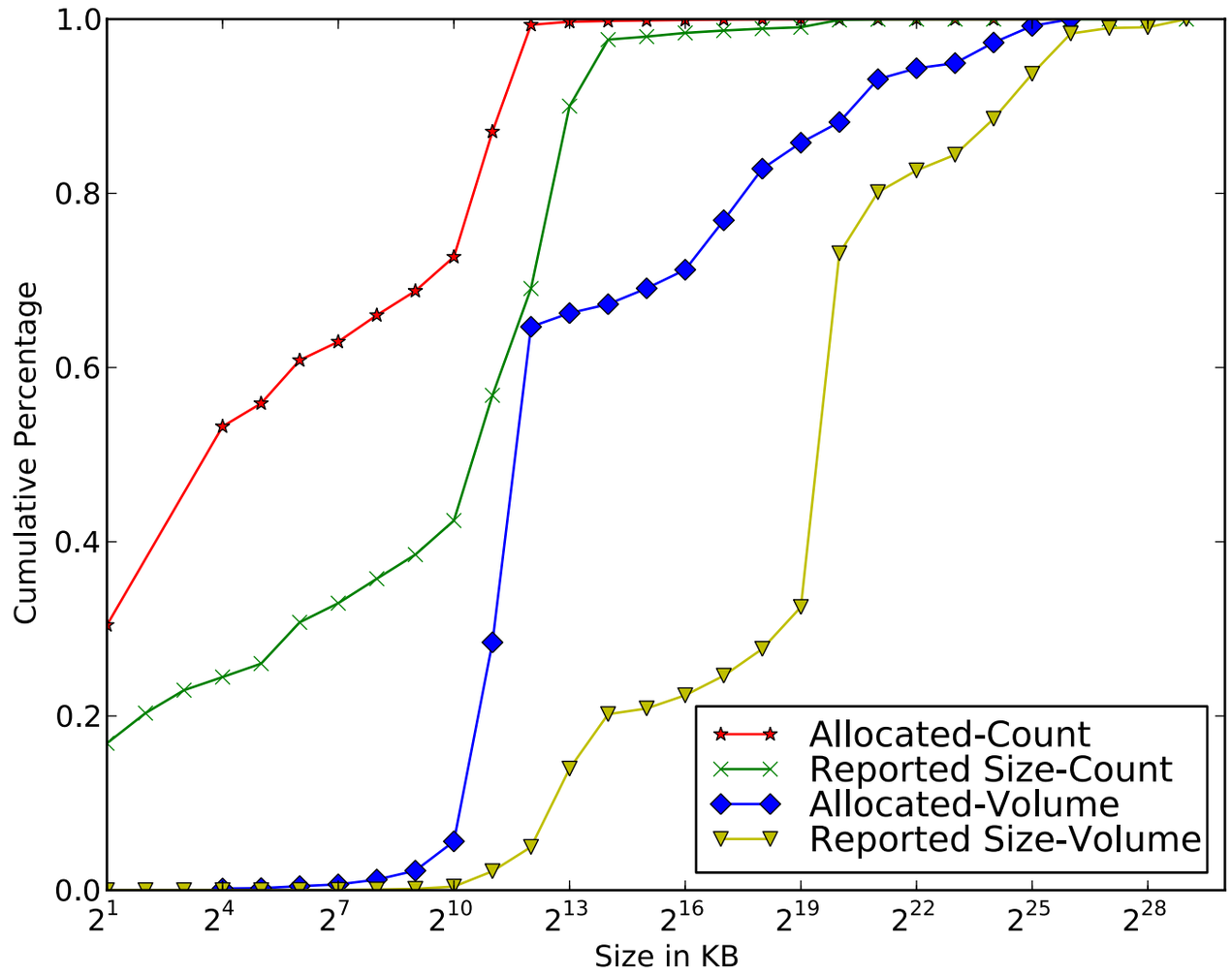
- Studying two HEC archives
  - Los Alamos National Laboratory (LANL)
  - National Center for Atmospheric Research (NCAR)

- LANL study uses periodic snapshots of statistics across high-level directories
- NCAR study uses access traces gathered over several years
  - Similar technique to earlier NCAR archive study from 1993 (!)
  - Still in early stages of the analysis

LANL study: file sizes

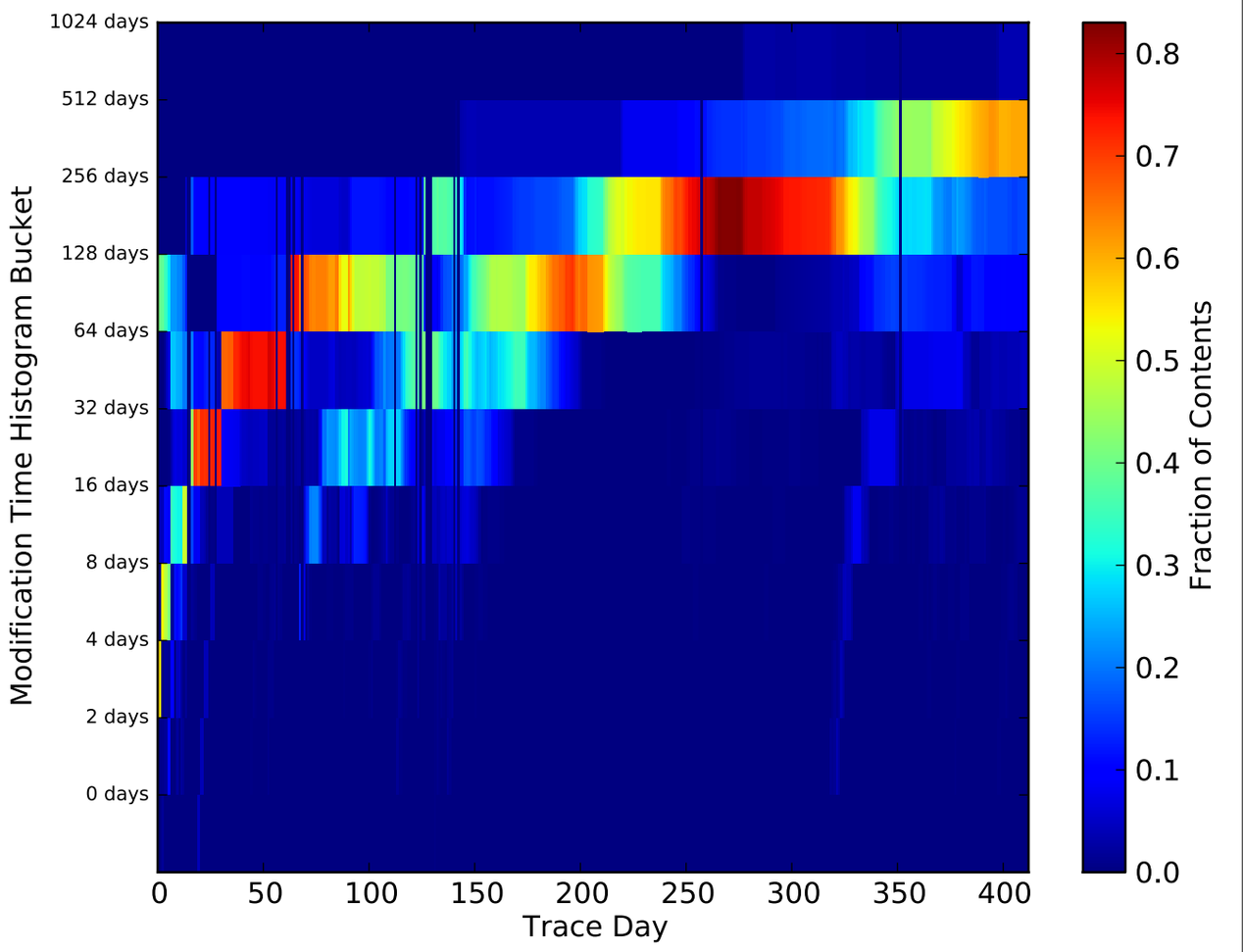
- Files 1–2 GB consumed 40% of reported space
- Many of these files were sparse
  - 60% of allocated space was consumed by files 2–8 MB

➡ Archives need to efficiently handle sparse files!

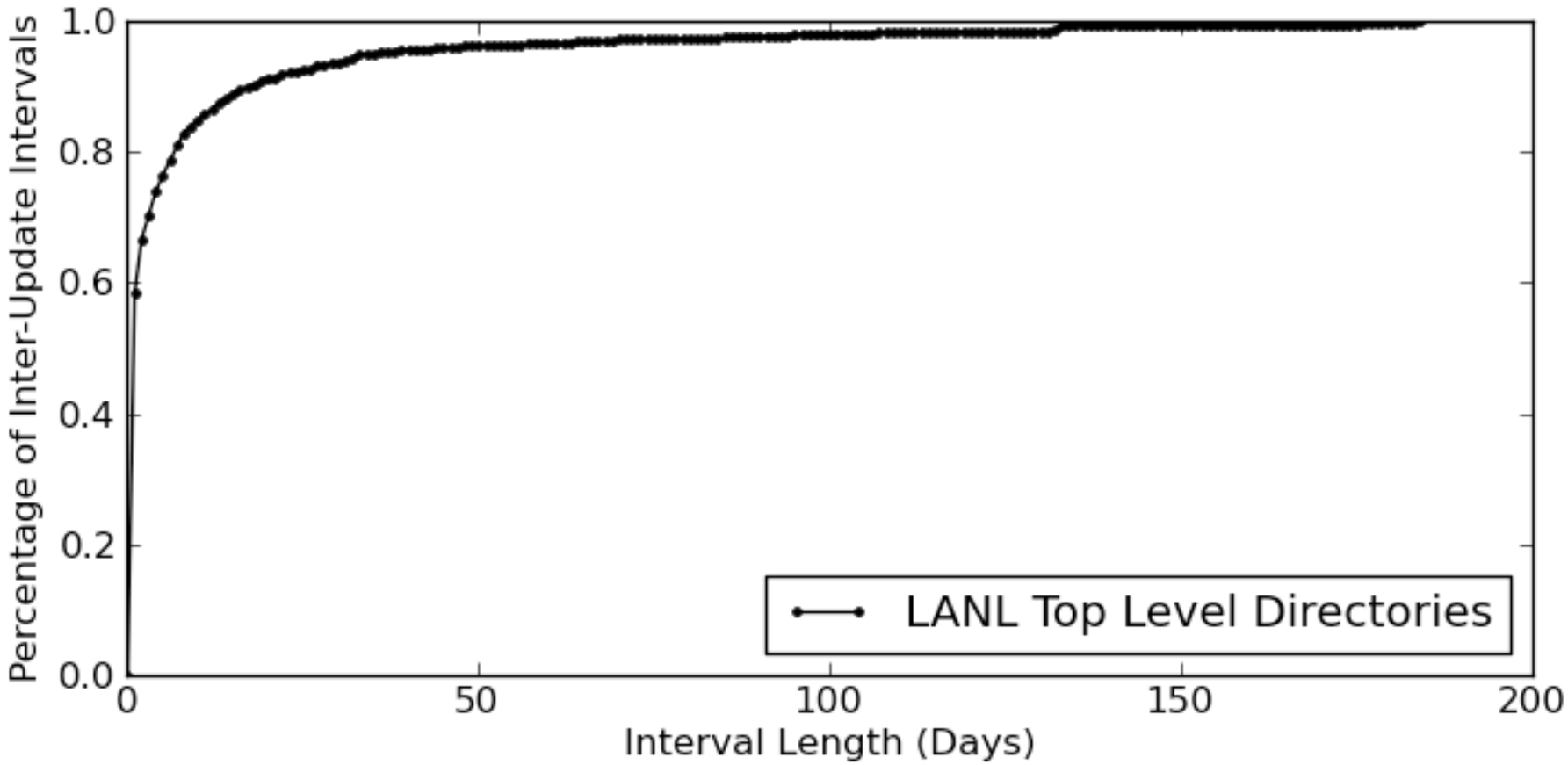


LANL study: system evolution over time

- Aggregate modification behavior similar to 1993 NCAR study
- System appears more disk-centric than prior studies
  - Have cheap disks shifted usage or caching behavior?



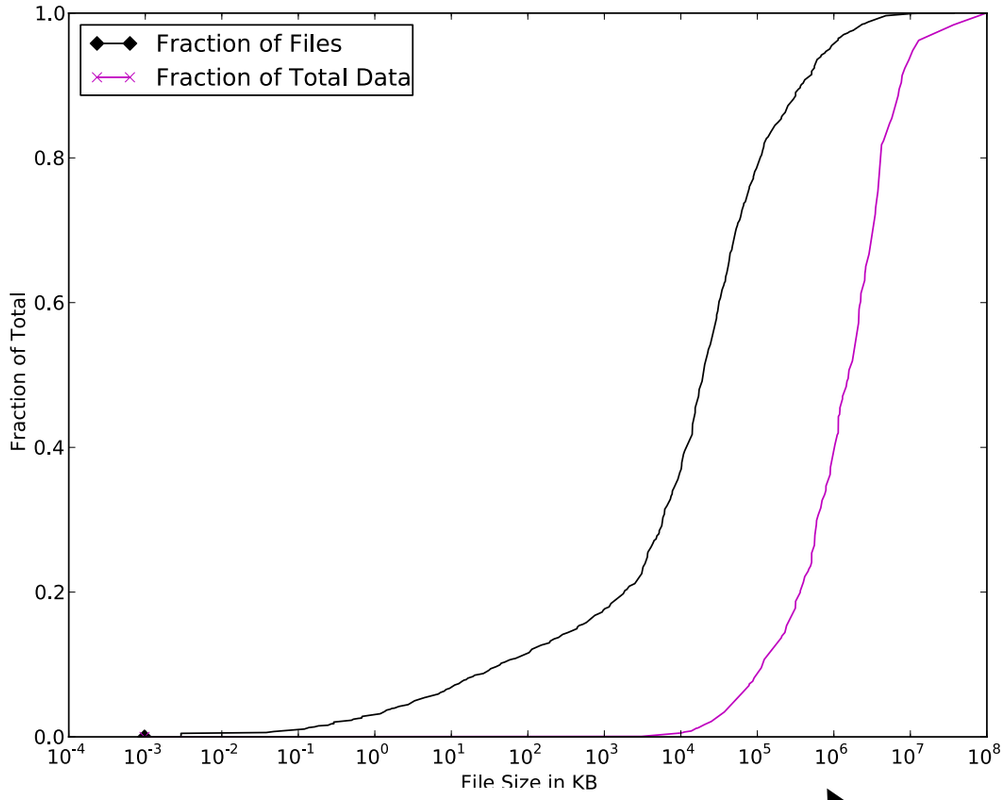
LANL study: directory inter-update interval

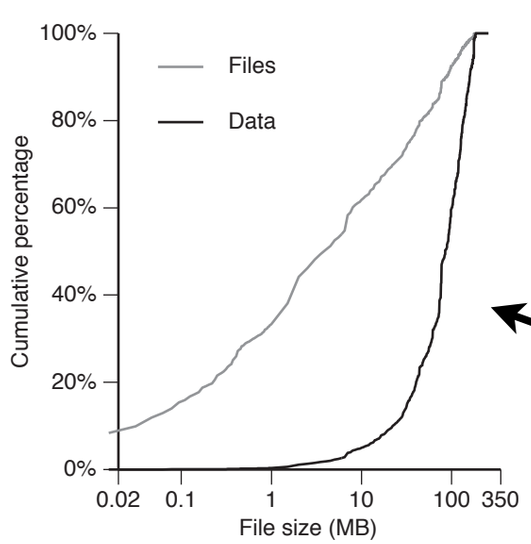


- Most updates to top-level directories came soon after the previous update
- Long tail: some updates were 100+ days after the previous update
- This has implications for caching and grouping

NCAR study: file sizes

- Most files are 10MB–1GB
  - 1–2 orders of magnitude larger than 1993
- Most data is in files 1–10GB
  - There were few files above 500MB in 1993
- Total storage has grown by 3+ orders of magnitude





2011

1993

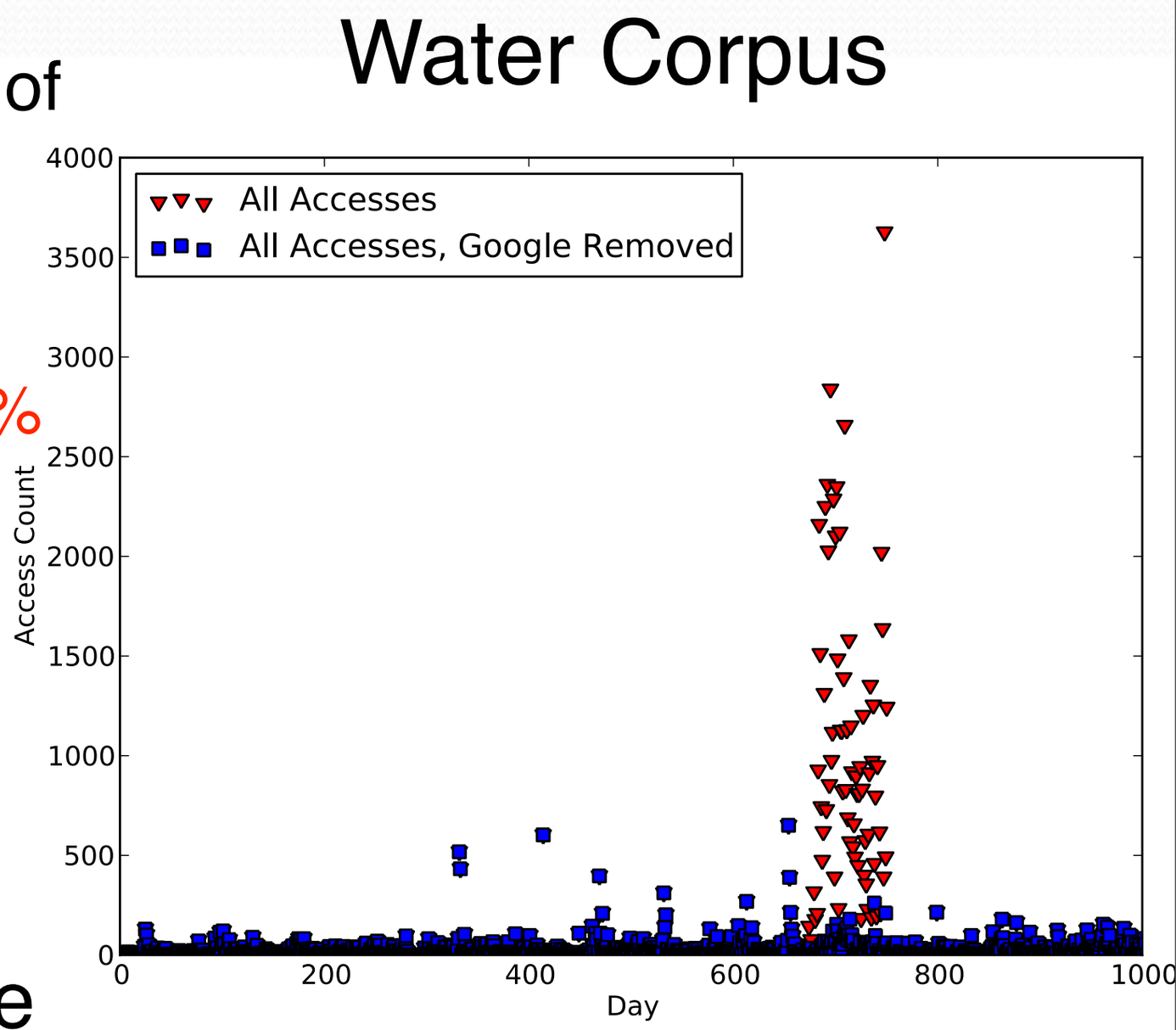
Other archival storage research

- Files in public archives are often not read-only
  - Artifact of updated results?
    - Example: multiple versions of a data set
    - Example: data set gathered over a long period of time
  - May need to have better support for writeable data
- Files in public archives are read more than previously thought
  - Google and other indexers
  - Batch interface for reading files would be very helpful
- Indexing and searching in archives is very important
  - Becoming even more critical as archives become larger and older

Public archive study: aggregate access patterns

- Mass accesses
  - Google accounts for 70% of water corpus retrievals
  - Integrity checking processes account for 99% of retrievals to historical corpus (not shown)
  - Updates to both corpora were done via batch processes

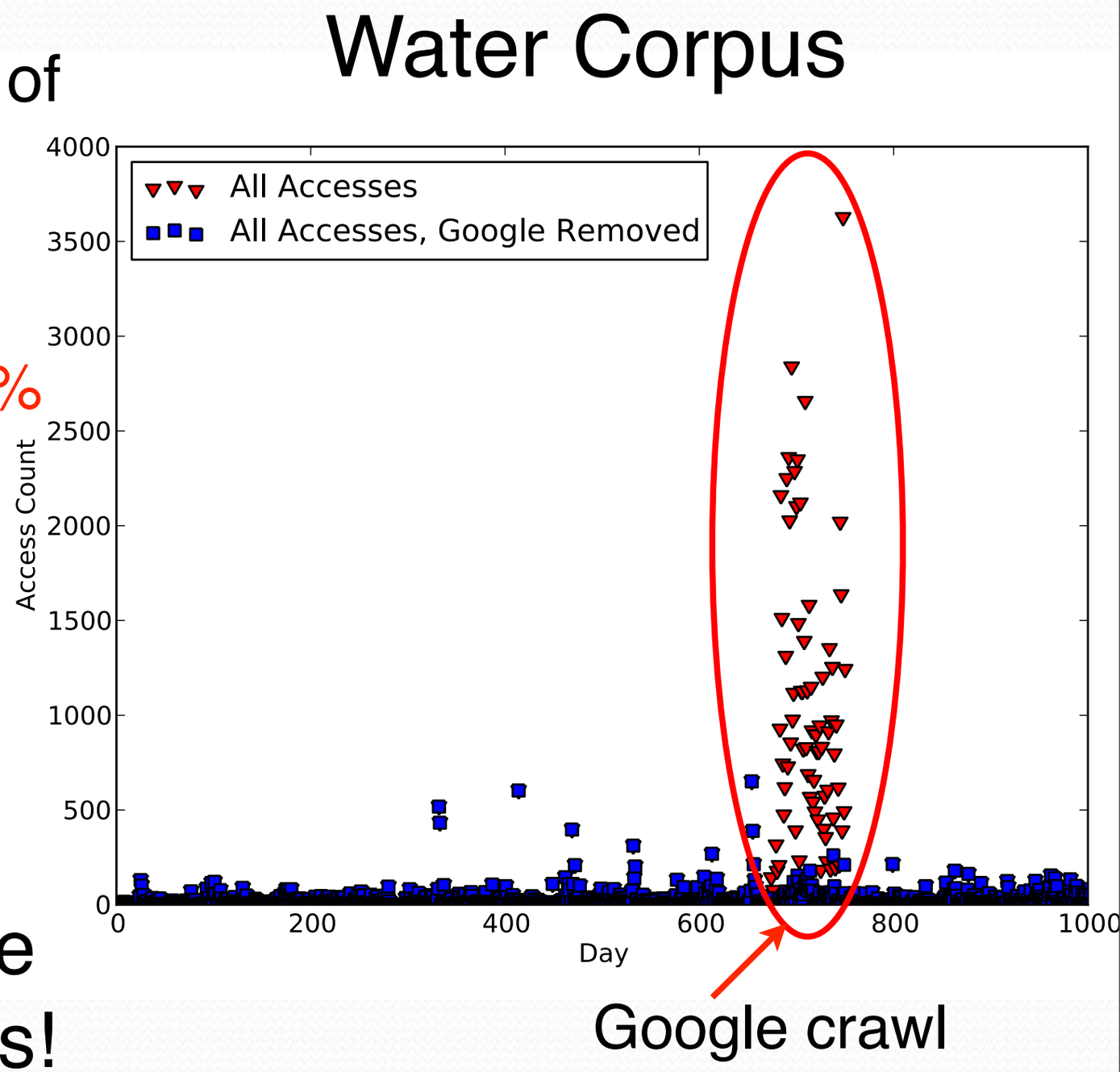
➡ Indexing & maintenance make up most accesses!



Public archive study: aggregate access patterns

- Mass accesses
  - Google accounts for 70% of water corpus retrievals
  - Integrity checking processes account for 99% of retrievals to historical corpus (not shown)
  - Updates to both corpora were done via batch processes

➡ Indexing & maintenance make up most accesses!



Conclusions

- Archival storage is of critical importance to the HEC community
  - Preserving data for the long term
  - Providing the ability to find and retrieve the data
  - Preserving the ability to *use* the data
- There's some research on how to do this, but not enough
  - It's a difficult problem!
  - It's a problem that requires long-term thinking
- Critical to solve the problem before scientific data becomes lost to the research community

Questions?


NSF


Office of Science
U.S. DEPARTMENT OF ENERGY


pdsi


iTRIS
Center for Information Technology
Research in the Interest of Society


EMC²


IBM


NORTHROP GRUMMAN


hp


NetApp™


SAMSUNG